Large Language Model Alignment: A Survey

Year: 2023 | Citations: 271 | Authors: Tianhao Shen, Renren Jin, Yufei Huang

---

Abstract

Recent years have witnessed remarkable progress made in large language models (LLMs). Such advancements, while garnering significant attention, have concurrently elicited various concerns. The potential of these models is undeniably vast; however, they may yield texts that are imprecise, misleading, or even detrimental. Consequently, it becomes paramount to employ alignment techniques to ensure these models to exhibit behaviors consistent with human values. This survey endeavors to furnish an extensive exploration of alignment methodologies designed for LLMs, in conjunction with the extant capability research in this domain. Adopting the lens of AI alignment, we categorize the prevailing methods and emergent proposals for the alignment of LLMs into outer and inner alignment. We also probe into salient issues including the models' interpretability, and potential vulnerabilities to adversarial attacks. To assess LLM alignment, we present a wide variety of benchmarks and evaluation methodologies. After discussing the state of alignment research for LLMs, we finally cast a vision toward the future, contemplating the promising avenues of research that lie ahead. Our aspiration for this survey extends beyond merely spurring research interests in this realm. We also envision bridging the gap between the AI alignment research community and the researchers engrossed in the capability exploration of LLMs for both capable and safe LLMs.